Identifying Key Players in Soccer Teams using Network Analysis and Pass Difficulty

Ian McHale* Samuel D. Relton†

November 22, 2017

Abstract

We use a unique dataset to identify the key members of a football team. The methodology uses a statistical model to determine the difficulty of a pass from one player to another, and combines this information with results from network analysis, to identify which players are pivotal to each team in the English Premier League during the 2012–13 season. We demonstrate the methodology by looking closely at one game, whilst also summarising player performance for each team over the entire season. The analysis is hoped to be of use to managers and coaches in identifying the best team lineup, and in the analysis of opposition teams to identify their key players.

Keywords: sport; big data; football; Moneyball; random effects

1 Introduction

The use of quantitative analysis in sports is, like in many industries, on the rise. A combination of increased computing power, and better recording and availability of data, has led to an increase in the awareness of the contribution that analytics can make to success in the sporting arena. Across sports globally, there are success stories to be cited. Perhaps the first, and certainly the most well-known case of analytics being used successfully in sports, is the story of how the Oakland Athletics were able to compete at the very highest echelons of Major League Baseball on a budget of around a tenth that of the bigger teams. Elsewhere, in cycling for example, much of the success of British riders in the Tour de France and at the Olympics Games in recent years has been attributed to Sir Dave Brailsford's adoption of analytical methods.

In soccer however, there is as at the time of writing, no such "success story". Further, there is very little written evidence documenting the adoption of advanced quantitative methods in the pursuit of gaining an advantage over the opposition by any professional team. The seemingly slow take up of analytics in soccer is very likely a consequence, at least to some extent, of the complexity of the game: 22 players moving and interacting continuously for over 90 minutes is certainly not a simple setting for an analyst, and makes it particularly difficult to gain insight above what an expert eye can achieve.

But recent advances in data collection has meant that rich, detailed data on the locations and timings of all actions on the pitch are now available. Such attractive data sets have caught the eye of academics, and in the academic literature there are now some examples of utilising such data. McHale and Szczepanski (2013) and Szczepanski and McHale (2015) present models for identifying goal scoring ability and pass making ability respectively. Meanwhile Peña and Touchette (2012) take an entirely novel approach to the analysis of football strategy and make use of network analysis to identify the important players on each team.

In addition to data detailing the location of events, the richest datasets available in soccer also give the locations of the 22 players themselves. Recorded at a frequency of up to ten times per second, the 'player tracking data' can be used to measure distance covered, top speed, and the acceleration of players. Indeed, to date, the majority of academic work using such data has been descriptive in nature.

*Corresponding author address: Centre for Sports Business, University of Liverpool Management School, University of Liverpool, UK. (ian.mchale@liverpool.ac.uk)

†Leeds Institute for Health Sciences, The University of Leeds, Leeds, LS2 9LU, UK. (s.d.relton@leeds.ac.uk)

For example, Castellano et al. (2014) assess the accuracy of two systems for tracking players on the pitch, whilst Rampinini et al. (2007) tabulate speeds and distances run by players and compare these between the first and second halves of matches.

Outside of soccer, player tracking data has been used by Cervone et al. (2014) to calculate the expected possession value in basketball. This methodology estimates the impact each player has on the probability that a series of passes (the possession) results in points being scored. The probability is updated continually as players move around the court and the players are 'rewarded' for their actions which contribute to increases in the expected value of the possession.

In this paper we also make use of player tracking data. Collected and made available to us by Prozone, we have information on the location (x-y coordinates) of each player at a frequency of ten times per second. The data also include the events occurring in the match (such as passes, tackles, dribbles and shots etc.). We have these data for all 380 matches in the 2012–13 season of the English Premier League season.

Our objective is to use this unique dataset to learn about which players are key to each team. Such analysis and information could be used by team managers and coaches to aid decision making in team selection, and where to concentrate effort on the pitch in order to thwart the opposition's strengths. Our model fills a gap in the literature since player tracking data have, until now, not been used in soccer in any meaningful way to inform team strategy or recruitment. Further, by utilising such rich data, the resulting tools we develop should be able to identify key players more accurately than previously available models.

To achieve our objective, we combine two tools: network analysis and statistical modelling. The use of network analysis is intended to identify the key passers in the team – those players which are heavily involved in passing moves, and who are central to how the team plays. However, to take account of the impact a player's passes have on the team, we weight the passes in terms of importance. We do not know the importance of the pass, but we proxy it using a measure of pass difficulty, which we take as the probability of the intended pass being successful. And this is our second tool, a statistical model to estimate the probability of a pass being successful.

We use this weighting scheme because it should, in principle, reflect players frequently involved in passing moves from which the ball enters key areas that are heavily defended by the opposition team. Thus the measure of pass difficulty, should in theory be related to pass importance. A player making 5 yard passes to the side, or even backwards, in his own half is likely to be much less effective in generating goal scoring opportunities than a player passing into the opposition penalty area. Such a weighting scheme should help identify players at the heart of a team's "shot generation engine", and knowing which players these are has clear advantages when selecting which players to field, and how to nullify the attacking threat posed by opposition teams.

The paper is structured as follows. First we present the data and give some descriptive statistics in section 2, before discussing our model for estimating the probability of a pass being successful in section 3. Section 4 presents the network analysis tools we employed to generate our results in section 5. We conclude with some closing remarks in Section 6.

2 Player Tracking Data

The depth of the analysis that can be performed to analyze player and team performance is enormously dependent upon the data that one has available. In soccer, the most widely available data are simple summary statistics for each game. For example, we might know that a player had a 95% pass completion rate in one game. However, without knowing the context of each pass, it is hard to judge whether or not this is an impressive feat, and as such, the insight that can be gleaned is massively diminished. For example, a midfielder making lots of passes back towards the defence will have a high completion rate but few of these passes would contribute towards winning the game. The rich nature of player tracking data makes much deeper analysis possible.

The data used in this research, provided by Prozone, gives the x-y coordinates of each player ten times per second to 10cm accuracy, and was made available to us for all 380 games in the English Premier League during the 2012–13 season, leading to a dataset containing over 451 million player positions and over 960,000 events. In the era of 'big data', this data must qualify. In the remainder of this section we present some descriptive statistics on this unique and rich dataset.

Table 1: Average distance run by each playing position during the English Premier League 2012–13. Running is defined as moving at more than 3 m/s, whilst a sprint is more than 6m/s.

Position	Total Distance (km)	Running Distance (km)	Sprints per 90 mins
Goal Keeper	5.4	0.5	0
Centreback	9.6	3.6	62
Fullback	9.9	4.1	84
Wide midfield	8.5	3.8	72
Centre Midfield	9.2	4.2	80
Attacker	7.6	3.1	64

Table 2: Average distance covered and other statistics by teams in the English Premier League 2012–13.

League Position	Team	Distance Covered (km)	Shots	Goals	Passes into final third	Number of sprints per 90 mins
1	Man. Utd.	8.6	560	86	3348	65
2	Man. City	8.6	659	66	2558	72
3	Chelsea	8.9	627	75	2957	67
4	Arsenal	8.5	598	72	2876	68
5	Tottenham Hotspur	8.5	681	66	2775	65
6	Everton	8.8	633	55	2035	69
7	Liverpool	8.9	740	71	4072	72
8	West Brom. Alb.	8.7	506	53	2229	76
9	Swansea City	8.6	506	47	3885	67
10	West Ham Utd.	8.5	493	45	1932	71
11	Norwich City	8.6	413	41	1896	71
12	Fulham	8.7	460	50	3298	70
13	Stoke City	8.7	390	34	1531	68
14	Southampton	9.2	516	49	2427	80
15	Aston Villa	8.6	438	47	2258	66
16	Newcastle Utd.	8.6	532	45	2765	69
17	Sunderland	8.8	417	41	1767	71
18	Wigan Athl.	8.7	500	47	3009	69
19	Reading	9.0	393	43	1662	72
20	Queens Park Rangers	8.3	500	30	1997	64

Table 1 shows the average distance covered by players, for each playing position. As one would expect, goalkeepers cover the least distance, though it is perhaps surprising to see that even they cover over 5km per game as they protect their 6 yard wide goal. The most ground covered is by fullbacks (left backs and right backs). In the modern game, this is again unsurprising as fullbacks are charged with both attacking and defending duties. Of the outfield positions, attackers cover the least distance.

Also shown in Table 1 are the running distances and number of sprints per 90 minutes. The story is similar to the total distance covered - goalkeepers run much less and do no sprints per 90 minutes, whilst fullbacks and centre midfielders do the most.

Table 2 shows the distances covered per 90 minutes by each team (per player) over the season, ranked by final league position. Also shown are the number of passes into the final third of the opposition pitch, the number of shots, and the number of goals. It is interesting to see how well the descriptive statistics 'explain' league position. A simple way to do this is to calculate Spearman's rank correlation, ρ, between each of the descriptive statistics and the league position. Unsurprisingly, goals have the strongest relationship with league position ($\rho = 0.85$), followed by shots ($\rho = 0.69$), then passes ($\rho = 0.43$), all of which are statistically significant. The intriguing result here is that distance covered and number of sprints have negative rank correlations with league position of -0.08 and -0.11 respectively, though not statistically significantly different from 0. It appears then, that successful football teams are doing something other than simply running more, or performing a higher frequency of sprints.

Table 3: Top 10 players in the English Premier League (2012–13) by number of passes per game (PPG) along with their pass completion percentage.

Player	PPG	Completion %
Mikel Arteta	75.7	94
Michael Carrick	73.7	90
Yaya Toure	73.7	89
Santi Cazorla	64	86
Steven Gerrard	62.4	86
Darren Fletcher	61.3	91
David Silva	61.2	85
Bacary Sagna	59.6	88
Angel Rangel	58.4	83
James McCarthy	57.7	89

These results are somewhat contradictory to the popularly held belief that running more than the opposition improves the team's chances of success, and to examine this relationship further, we performed the following experiment. For the moments in each match when the scores were level (e.g. $0-0$, $1-1$, $2-2$ and so on), we calculated the total distance covered and the number of sprints by each team. We then used the difference between the home and away teams' distances covered per minute and numbers of sprints per minute as two covariates in a logistic regression model to predict whether the home team scored the next goal. We chose to use only moments in the match when the scores were level to guard against any bias that might be introduced since once a team goes ahead (behind) in a match, it is likely to change its behaviour. The reason for the bias is because it is not random which team takes the lead - it is more likely that the better team scores first.

The results of the model concur with the above result - the coefficients on *difference in distance covered per minute* and *difference in sprints per minute* are both statistically significant and negative. In other words, if the home team runs more than the away team, the probability that it scores the next goal is lower. As a check of the results, we refitted the model with *difference in passes per minute* and *difference in passes in the final third per minute*, and the estimated coefficients were statistically significant and positive, such that when the home team passes more than the away team it is more likely to score the next goal. We should note here that these results do not suggest that running less than the opposition will result in more success! Rather, the result is a consequence of the playing style of the better teams, and more skillfull players, in the league.

In this research we want to identify the key passers on each team. For comparison, in Table 3 we show the top 10 players in the league over the 2012-13 season, measured by their average number of passes per game, along with their pass completion rate.

3 A random effects model for passing difficulty

Having explored the data, we now move to the main objective of the paper - identifying the key passers on each team. Our first task in achieving this objective is to create a model that captures the probability of a pass being successful. The resulting estimated probability will be used as a proxy for pass importance, and will subsequently be used to weight the edges in our network analysis.

We use a generalised additive mixed model (GAMM), see, for example, Hastie and Tibshirani (1990), to estimate the probability of a pass being successful. Each pass is treated as a Bernoulli trial with the probability of a pass being successful depending on covariates. However, the advantage of using an additive model is that, unlike for linear models, smooth functions of the covariates can be included. Here we use a tensor product smooth function for the x-y coordinates of the origin and destination of the pass.

Let us denote the outcome of the ith pass by o_i where $o_i = 1$ is a successful pass and $o_i = 0$ is a failed pass. We assume that the distribution of each pass follows a Bernoulli distribution with the probability of success represented by the inverse logit function of the linear predictor η_i. That is we have

$$(o_i|\eta_i) \sim \text{Bernoulli}(p_i), \tag{1}$$

where

$$p_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}. \tag{2}$$

We let the linear predictor η_i be a function of covariates, such that

$$\eta_i = \mathbf{W_i}\beta + \mathbf{Z_i}\mathbf{b} + s(x_i, x_{i,end}, y_i, y_{i,end}), \tag{3}$$

where $\mathbf{W_i}$ is a row vector of covariates, $\mathbf{Z_i}$ is a row of a design matrix selecting the elements of the random-effects vector $\mathbf{b}$ corresponding to the player executing the ith pass, and s is a tensor product smooth function that we use to account for the origin and destination of each pass. This allows the probability of each pass being successful to be a function of the x and y coordinates of the origin and destination. Adopting a tensor product smooth here is much like using a spline in several dimensions, and the advantage is that no functional form need be specified for the shape of the relationship between the dependent variable and the 4-dimensional set of covariates (the $x - y$ coordinates of the pass origin and destination).

The inclusion of the random effect term $\mathbf{b}$, accounts for the fact that different players have different levels of skill. Szczepanski and McHale (2015) fit a similar model to this and used the estimated random effect term for each player as a way to rate players' passing abilities. Here we are not interested in the value of the random effect term itself. Rather, we use the model to estimate the probability of any given pass being complete, and the inclusion of the random effects term is necessary to avoid inducing a bias in the model results that would results from different players performing the passes.

To calculate the probability of any one pass being completed, we set the value of the random effect back to 0 (and ignore the identity of the player performing the pass). This is effectively like assuming the pass is being performed by a player with average passing ability. We do this because we do not want the player's skill to be included in the calculation of the pass difficulty measure. If it was, players with high passing ability would be penalised (their passes would be estimated to have a higher probability of being completed than they would otherwise be – i.e. an artificially low pass difficulty), whilst players with low passing ability would benefit (their passes would be estimated to have a lower probability of being completed – i.e. an artificially high pass difficulty).

A major advantage of our work over Szczepanksi and McHale (2015) is that we have access to player tracking data, so that we can create covariates that better account for the subtleties affecting the success of each pass. For example, we can "see" if the passing player or the receiving player is under pressure from opposition players; or we can measure how fast the passing player is moving with the ball. The full list of covariates we derived from the player tracking data, and their definitions, are as follows.

1. *dist:* the distance (in metres) from the origin of the pass to the intended destination. We expect this to be negatively correlated with success.

2. *moe (margin of error):* the minimum angle from the line between the pass origin and destination to any opposition player (see Figure 1). We expect this to be positively correlated with success.

3. *intendedX:* the intended x coordinate of the pass.

4. *intendedY:* the intended y coordinate of the pass.

5. *forwardPass:* a dummy variable indicating that the ball was intended to move closer to the opposition goal line.

6. *aveXoppTeam:* the average distance on the x-axis from the opposition goal of the opposition team players. This variable attempts to capture a counter-attack from the attacking team, in which case they may have more room to pass the ball.

7. *aveXpassersTeam:* the average distance on the x-axis from the opposition goal of the passing players team. This is intended to capture a counter-attack in conjunction with *aveXoppTeam*.

8. *passerPressure:* the number of opposition players within a 4 metre radius of the passer.

9. *receiverPressure:* the number of opposition players within a 4 metre radius of the intended receiver.

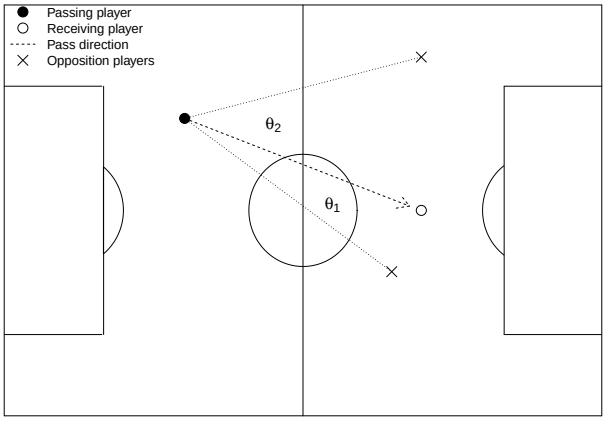

Figure 1: Margin of error. The solid circle represents the player with the ball and the hollow circle the teammate he is intending to pass to. The opposition players are shown as crosses. θ_1 is the margin of error as it is the minimum angle between the intended direction of the pass and an opposition player.

10. *touches* and *firstTimePass*: the number of touches the passing player has had, and a dummy variable indicating that the passer played the pass with no previous touches of the ball. We would expect first touch passes to be more difficult.

11. *timeOnBall:* the time on the ball the passing player has had.

12. *home:* a dummy variable indicating that the passer is playing at their home ground. There may be some effect on the pass success probability based on familiarity with the surroundings.

13. *passersTeam:* a variable to account for the quality of the passer's team in being able to receive the ball.

14. *oppositionTeam:* a variable to account for the quality of the opposition in stopping passes being completed.

15. *time:* the minute of the match which may pick up tiring of players and a subsequent change in the probability of a given pass being completed.

16. *passNumberInPossession:* a variable counting the pass number in the current possession.

17. *previousPassSuccess:* a dummy variable indicating whether the previous pass in the possession sequence was successful.

We fitted this GAMM with the use of the R package `gamm4` by Wood and Scheipl (2016). We used 90% of the data as a training set for the model and looked for evidence of overfitting on the remaining 10% of the data. The Brier Scores achieved in both sets of data were similar suggesting no evidence of overfitting.

Table 4 shows the estimated values of the model coefficients (the vector β in equation (3)). These are the final set of covariates included in the model. Other covariates mentioned in the list above that are not included in the model were not statistically significant. Before discussing the covariates that were included in the final model, it is worth commenting on some of the covariates that have been omitted. First, unlike in Szczepanski and McHale (2015), time in the match was not significant. We think this is because McHale and Szczepanski did not have information on the location of the opposition players. As such, the probability of a pass being completed increased as the match progressed perhaps because the opposition players did not surround the passer and receiver as much as they did in the early part of the match. In our model, this is controlled for through the *passerPressure* and *receiverPressure* variables. Second, we investigated whether the outcome of consecutive passes were independent using two variables: *passNumberInPossession* and *previousPassSuccess*. Neither of these variables were statistically

Table 4: Parameter estimates for the fitted Binomial Additive Model. We show the estimated values of the covariates derived from the Prozone data. Standard errors are shown in parentheses.

Parameter	Estimate (s.e.)	z-value
Intercept	1.516 (0.05)	27.87
dist	-0.872 (0.01)	-115.78
moe	4.780 (0.07)	70.17
moe^2	-1.654 (0.05)	-36.52
aveXoppTeam-aveXpassersTeam	-0.010 (0.00)	-3.12
passerPressure	-0.201 (0.05)	-3.90
receiverPressure	-0.189 (0.05)	-3.52
timeOnBall	0.019 (0.02)	1.24
timeOnBall^2	-0.006 (0.00)	-3.32
forwardPass	-0.289 (0.03)	-11.29
firstTimePass	-0.595 (0.03)	-21.60

significant, again suggesting that the *passerPressure* and *receiverPressure* variables are taking account of any possible impact of consecutive passes.

Examining the values and signs of the estimated coefficients on the covariates included in our model in Table 4 reveals factors that affect the difficultly (or ease) of a pass. The smooth tensor term for the origin and destination of the passes is strongly statistically significant, and, as we might expect, the intended distance of the pass, the pressure on the passer and receiver, and the distance of the passer's team from the opposition goal are all negatively related with pass success. First time passes are particularly difficult. In fact, the relative effect of a first time pass is similar to having three players within 4m (*passerPressure*) of the passer. Meanwhile the margin of error, the time on the ball, and the distance of the opposition team from their goal line are all positively related with success. From this it is clear that a player with more room and more time to assess their position has a better chance of completing a pass.

3.1 Model Diagnostics

In this section we examine how accurately the model predicts pass success. Standard diagnostic plots revealed that the model assumptions were satisfied. A simple gauge of how the model is performing is to compare its Brier Score with that of a baseline model. We choose the baseline model to be the grand mean pass success rate of 85.7%. Taking this as the predicted pass success probability results in a Brier Score of 0.123. The predicted probabilities from the model result in a Brier Score of 0.079 demonstrating a considerable improvement in goodness-of-fit.

Next we look at the calibration of the model. A perfectly calibrated model *knows* how often it is wrong. For example, if an event is predicted to happen with a probability of 70%, then the event should occur 70% of the time. Figure 2 shows the 'calibration curve' for the passing model. The x-axis shows the model's predicted probability of a pass success, and the y-axis shows the observed frequency of pass success. A perfectly calibrated model would have points lying on the $y = x$ line, such that the model probability equalled the empirical frequency. Here each point represents many passes, where passes are binned according to Tukey's (Tukey, 1961) approach of dividing the prediction space by 'halves'. For more detail, see Boshnakov et al. (2017). The size of each point is made proportional to the number of passes it represents. It is encouraging to see that all of the points lie on, or very near, the $y = x$ line suggesting that the model is indeed 'well-calibrated'.

In Figure 3 we show heatmaps comparing the predicted probability of a pass being successful with the actual outcome. In both figures the passer's team are playing left-to-right. The left hand plot shows, for each intended pass destination coordinate, a heatmap of the predicted probability of a pass being successful. The right hand plot shows a heatmap of the model error (actual outcome minus model probability). From the first heatmap we see that our model predicts that passes are generally completed with high probability everywhere except for near the opposition goal, and that passes with destinations inside the opposition's penalty box are more difficult.

The second heatmap reveals that outside the opposition's penalty box, the model has very low error, including on the passing team's penalty spot (which has a lower probability of pass success than other

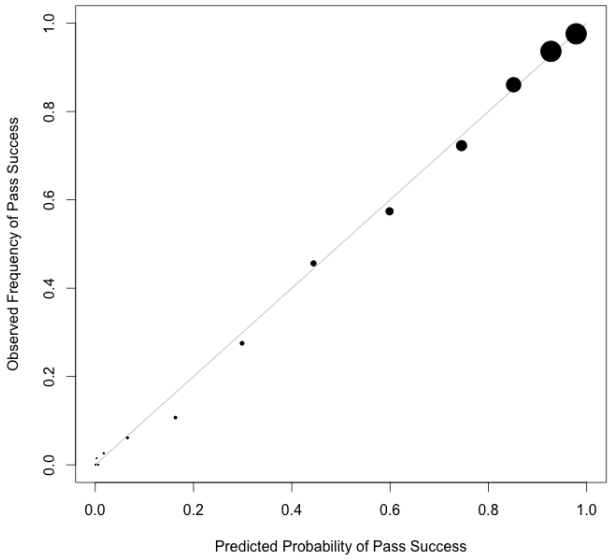

Figure 2: Calibration curves for predicting the outcome of a pass. The size of the circles are proportional to the number of observations in each bin.

areas). Within the opponent's penalty area we see both under- and over-estimation, suggesting that the model is not biased in this region of the pitch and the inaccuracy is due to high volatility. The particularly dark spot on the bottom-edge of the opponent's penalty area is due to there being only one pass in our dataset being destined for this area, which the model misclassified (which is also why this region has 0 variance in Figure 4).

Note that from this representation it is not possible to see the effect that the other variables such as *dist* and *moe* have on the pass success rate.

To further examine the cause of the model error in the opposition penalty box we plot the variance of pass success in Figure 4. From this heatmap it is clear that the opposition penalty box is a highly volatile area, in comparison to the rest of the pitch, and it may therefore be difficult for any model to accurately predict pass success in this region. To corroborate this, we also fitted a random forest classifier to the data and obtained similar model errors within this area. But of course, the advantage of using a regression approach is the interpretability of the estimated coefficients.

4 Network analysis

In this section we explain how our passing model can be combined with techniques from network analysis to analyze the importance of each player within the team. We begin by recalling the basic principles of network analysis required, before explaining how to incorporate the pass success model from the previous section.

A network is formed by a number of *nodes* which are linked by *edges*. In our case, each player will be a node and the nodes are linked by passes to one another. In the simple case where all edges have the same value, we can represent this as an *adjacency matrix*, A, where node i and j are linked if $A_{ij} = 1$ and $A_{ij} = 0$ otherwise. This is called an *unweighted* network, and means that each team will have an associated adjacency matrix of size 11×11 for each game, if there are no substitutions during the match. Alternatively, we can consider a *weighted* network where each edge has an associated strength and A_{ij} is set to some real number.

There are two quantities that are particularly interesting for us: the *centrality* of a node and the *betweenness* of a pair of nodes. The centrality measures the importance of each node in the network (to identify the key players) whilst the betweenness measures the strength of the connection from one node

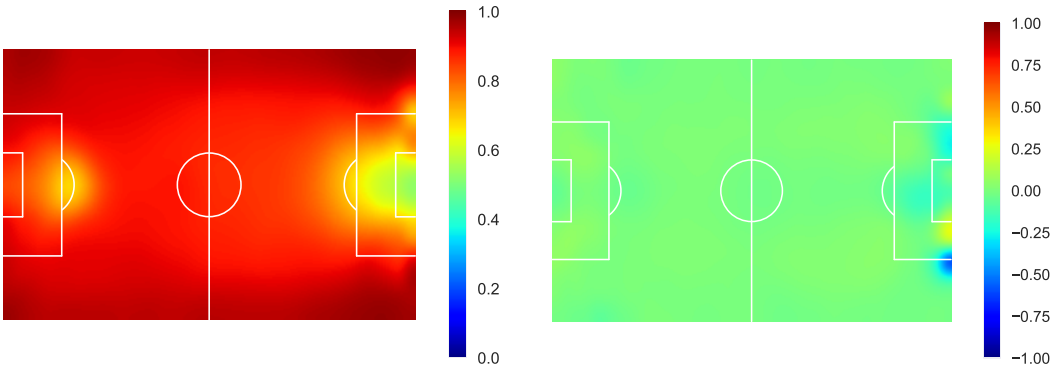

Figure 3: Heatmaps of pass success probability (left), and model error (right) by the intended pass destination. Model error is defined as $o - p$ where o is the binary outcome of the pass and p is the probability assigned by our model.

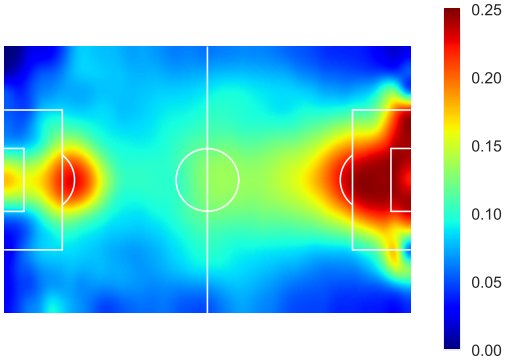

Figure 4: Variance of pass success probability by the intended pass destination.

to another.

There are several different ways to measure centrality although, a priori, there is no objective way to decide which will perform best on any individual problem. This is particularly true in our case since the existing rankings of players are inherently subjective, meaning there is no ground truth we can aim to emulate. Following some experimentation, we chose to use the exponential centrality defined by Estrada and Hatano (2008) as opposed to, for example, Katz centrality. This choice was partly to avoid the additional parameter needed for Katz centrality and partly because the resulting graphs and rankings were subjectively more in line with our intuition.

For node i the exponential centrality is defined as

$$C(i) = \exp(A)_{ii}, \tag{4}$$

and the exponential centrality betweenness for a pair of nodes (i, j) is defined as

$$B(i, j) = \exp(A)_{ij}, \tag{5}$$

where $\exp(A) = \sum_{k=0}^{\infty} A^k/k!$ is the matrix exponential. Note that our adjancency matrices are directed (i.e. nonsymmetric) and therefore $B(i, j) \neq B(j, i)$. We will therefore define the strength of the connection between nodes i and j as

$$S(i, j) = \frac{B(i, j) + B(j, i)}{2}. \tag{6}$$

Now that the basic concepts have been established we can combine the network analysis with our pass probability model. The motivation for doing this is that not all passes are equally important: two defenders passing back and forth to one another is less important than a midfielder making a difficult

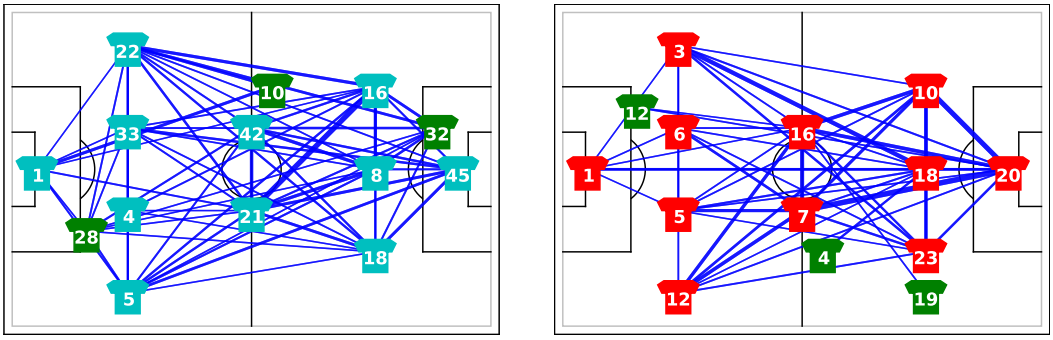

Figure 5: Passing networks for Manchester City (left) and Manchester United (right). Thickness of the lines is correlated with the betweenness of each pair of nodes.

pass to a striker. As such, if player i passes to player j successfully with a probability of success p we want to update A_{ij} by applying a weight to it which we denote by $f(p)$, where p is the probability of the pass being successful, and f is some function which converts the probability into a "difficulty rating". For example, we could use $f(p) = 1/p$ or $f(p) = 1 - p$ and the function f controls how much relative value is assigned to easy and difficult passes. For example, using $f(p) = 1/p^3$ would severely punish "easy" passes compared with more difficult ones.

We experimented with various specifications for f and in terms of the resulting ranking of players, using any monotonically decreasing function f appeared to make very little difference to the identity of the top players. The main differences were in the interpretability of the network plots (see next section), since the thickness of the lines connecting two players is relative to the difficulty function. After experimenting with several such functions f, we decided on using $f(p) = 1 - p$. This choice provides a clear distinction between weak and strong links in our plots, meaning that it is easier to draw insight from the visualisations.

5 Results

In this section we show how our techniques can be used to assess player importance in a team. We begin by presenting an in-depth analysis of a particular game between two of the English Premier League's biggest teams, and biggest rivals: Manchester United and Manchester City. Our technique picks out some interesting features of the game.

Following this in-depth game analysis, we create one large network per team for the entire season and use the exponential centrality measure to determine the key players in each team.

5.1 In-depth game analysis

Here we analyse the game between Manchester City and Manchester United on 9th December 2012, where Manchester United, playing away from home, won 3–2.

In Figure 5 we show two networks, one each for each team. The thickness of the lines between each pair of players increases with the connection strength (6) so, using Manchester United (on the right) for example, the connection strength between 10–Wayne Rooney and 20–Robin van Persie is much stronger than the connection between 23–Tom Cleverley and 20–Robin van Persie.

A coach or analyst can note a number of things from these networks. For instance within Manchester City 21–David Silva has a stronger connection with 16–Sergio Aguero than any of the other players on their team. On the other hand, 18–Gareth Barry, has the strongest connection with the other striker 45–Mario Balotelli. Looking at Manchester United we can see that 7–Antonio Valencia has a strong connection to 20–Robin van Persie - something future opposition teams could have acted upon.

In addition to examining network graphs, we can also look at the exponential centrality of each player, defined in equation (4), to determine which are the most important nodes (players) in the network (team). Table 5 shows the rankings for each player on each team. Such an analysis may be interesting for planning

Table 5: Importance ranking of Manchester City and Manchester United players by exponential centrality. Also shown are the approximate playing positions for this game. The abbreviations are as follows: GK=goalkeeper, CB=centre back, LB=left back, RB=right back, CM=centre midfield, LM=left midfield, RM=right midfield, ST=striker. Any position preceded by "sub" means that player made an appearance as a substitute.

Rating	Manchester City (playing position)	Manchester Utd. (playing position)
1	21 David Silva (CM)	7 Antonio Valencia (RM)
2	16 Sergio Aguero (ST)	20 Robin van Persie (ST)
3	8 Samir Nasri (RM)	10 Wayne Rooney (ST)
4	42 Yaya Toure (CM)	16 Michael Carrick (CM)
5	45 Mario Balotelli (ST)	23 Tom Cleverley (CM)
6	32 Carlos Tevez (sub: ST)	12 Rafael (RB)
7	22 Gael Clichy (LB)	18 Ashley Young (LM)
8	5 Pablo Zabaleta (RB)	3 Patrice Evra (LB)
9	18 Gareth Barry (LM)	5 Rio Ferdinand (CB)
10	33 Matija Nastasic (CB)	6 Jonny Evans (CB)
11	28 Kolo Toure (CM)	1 David de Gea (GK)
12	1 Joe Hart (GK)	12 Chris Smalling (sub: CB)
13	4 Vincent Kompany (CB)	4 Phil Jones (sub: CM)
14	10 Edin Dzeko (sub: ST)	19 Danny Welbeck (sub: ST)

Table 6: Top 5 importance ranking of Manchester City and Manchester United players by exponential centrality using an *unweighted* adjacency matrix. Also shown are the approximate playing positions for this game. The abbreviations are as follows: GK=goalkeeper, CB=centre back, LB=left back, RB=right back, CM=centre midfield, LM=left midfield, RM=right midfield, ST=striker. Any position preceded by "sub" means that player made an appearance as a substitute.

Rating	Manchester City (playing position)	Manchester Utd. (playing position)
1	21 David Silva (CM)	16 Michael Carrick (CM)
2	42 Yaya Toure (CM)	7 Antonio Valencia (RM)
3	18 Gareth Barry (LM)	10 Wayne Rooney (ST)
4	8 Samir Nasri (RM)	23 Tom Cleverley (CM)
5	22 Gael Clichy (LB)	12 Rafael (RB)

the strategy of an upcoming game: we observe that 21–David Silva is a key player in the Manchester City squad so it makes sense for their upcoming opponents to mark him closely, and try to neutralise his contribution to Manchester City's play.

As we might expect towards the bottom of the ranking are the goalkeepers and substitutes. The goalkeepers do not interact with the ball as much as the other players and, when they do, they typically attempt simple passes to defenders in close proximity. Similarly, substitutes have less time to make passes. The exception to this is 32–Carlos Tevez who, despite replacing 45–Mario Balotelli in the 52nd minute, ranked in 6th place for Manchester City.

The top of the table is primarily occupied by midfielders and strikers. This is expected for two reasons. First, the midfielders move the ball between the defence and strikers and typically perform more passes in a game. Secondly, defenders often perform passes that are simple and such "easy" passes are not highly rated in this model once we convert from pass probability to pass difficulty (see section 4).

To show the difference that incorporating our difficulty rating can have on the results of a network analysis in football, we compare our results to the methodology of Peña and Touchette (2012). The top five ranked players for each team using the *unweighted* adjacency matrix of Peña and Touchette are shown in Table 6. For Manchester City we see that 18–Gareth Barry and 22–Gael Clichy have replaced 16–Sergio Aguero and 45–Mario Balotelli. Since 22–Gael Clichy is a defender and 18–Gareth Barry is ranked 9th in Table 5 but is now 3rd, we can imagine that they completed more "easy" passes, which

are treated equally to difficult passes in the unweighted model. Indeed 18–Gareth Barry had a mean pass success probability of 0.78 and 22-Gael Clichy of 0.76 whilst 45–Mario Balotelli had a mean pass success probability of 0.73 corroborating our intuition. Looking at Manchester United we see that whilst 20–Robin van Persie has been removed from the top five the others have merely shuffled. A closer look at the data shows that 20–Robin van Persie successfully made a few quite difficult passes which were highly rewarded in the weighted analysis shown in Table 5.

5.2 Network analysis over a season

In the previous section we used networks to analyse the performance over a single game, however in this section we create a network for an entire season. In doing so, we can identify the key players in the network over a longer time frame. Since not every player will play in every match, it is important to modify the adjacency matrix of the network to take this into account. The procedure we use to perform this modification is as follows: if players i and j were in k matches together then we divide both A_{ij} and A_{ji} by k before computing the exponential centrality. This is equivalent to having each player play in one full match with each teammate.

The results for five of the top teams in the league are shown in Table 7. In addition to the identities of the three most 'central' players to the teams, the pass completion percentage and average probability of the player's pass being completed (our proxy of pass difficulty) are also shown. In hindsight, we believe that football experts would struggle to disagree with these findings. For example, knowing that Luis Suarez and Gareth Bale were the most important players on the Liverpool and Tottenham sides during the 2012–13 season is not surprising. And to some extent, this validates our approach. It is interesting to note how the key players within each team can have significantly varyied average pass difficulty. For example, Mikel Arteta at Arsenal attempts the simplest passes of all players in our list, but is more central to his team than Aaron Ramsey who attempts more difficult passes.

Table 7: Season rankings of the top 3 players for a variety of teams.

Rating	Player	Team	Pass Completion %	Pass Difficulty
1	Santi Cazorla	Arsenal	86	0.18
2	Mikel Arteta	Arsenal	94	0.10
3	Aaron Ramsey	Arsenal	83	0.14
1	Luis Suarez	Liverpool	76	0.21
2	Steven Gerrard	Liverpool	86	0.15
3	Glen Johnson	Liverpool	81	0.19
1	Eden Hazard	Chelsea	84	0.19
2	Juan Manuel Mata	Chelsea	82	0.18
3	Oscar	Chelsea	82	0.20
1	Gareth Bale	Tottenham Hotspur	74	0.24
2	Moussa Dembele	Tottenham Hotspur	94	0.11
3	Emmanuel Adebayor	Tottenham Hotspur	85	0.13
1	Michael Carrick	Manchester United	90	0.14
2	Wayne Rooney	Manchester United	83	0.16
3	Darren Fletcher	Manchester United	91	0.11

In the above analysis no account was taken account of the number of minutes played by players. This could result in substitute players getting penalised for playing fewer minutes. As such, we experimented with normalising the adjacency matrix by the number of minutes played by each player. Interestingly, the rankings remained very similar which we believe suggests that substitute players (players playing fewer minutes over the course of the season), are not as good as players playing more minutes.

6 Conclusions

We have presented a methodology for identifying the key players in a football team. The methodology utilises a unique and vast dataset which details the locations of all 22 players on the pitch at a frequency

of ten times per second. Examination of some descriptive statistics suggests that simply running more than the opposition is not necessarily positively related to team success.

The two key components of our approach to identifying key players are a statistical model to determine the probability of a pass being successful, and network centrality measures. For the former, we have shown that a generalised additive mixed model can accurately predict the probability of pass success in most areas of the pitch; whilst finding high levels of volatility in the opposition penalty box. This was coupled with the use of exponential centrality measures to identify key passers within a team.

By performing an in-depth analysis of a single game and summarising passing performance over a season, we have shown the utility of our approach. Our results also correlate well with expert opinion of player performance. For example, two of the most highly rated players in our model, Luis Suarez and Gareth Bale, were eventually bought for record transfer fees by Barcelona and Real Madrid in Spain.

The obvious use for our model is to help team owners and coaches identify playing talent (with a view to recruiting players), and for coaches to identify key passers on opposition sides. One can imagine a coach asking the players to be aware of, and to try to nullify the impact of, the key passers on the opposition side. It would even be possible to apply this model in real-time so as to identify the key passers and relationships on an opposition team in a specific game, perhaps at half-time. Further, one can use this model to identify key relationships between players; the example above of David Silva and Sergio Aguero would be valuable to opposition players.

The model adopted here could be extended in several ways. First, one could use it to identify players who perform well under different circumstances. For example, some players may maintain a high pass rating when their team is losing, whilst others may experience a deterioration in performance. Accounting for match situation like this could help coaches identify players who can perform when the pressure is greatest. Second, one may wish to modify the network definition to include unsuccessful passes. This would potentially differentiate between players perform many key passes with a low error rate (they do not lose possession), and players who perform an equally high number of key passes but at a higher error rate (and do lose possession a lot).

In future work we would like to use the data to further investigate the relationship between running and success which we alluded to in section 2. This could be approached by determining the success of different styles of play against one another, which is certainly related to the running statistics gathered here. For example, one might compare the effectiveness of a tactic with low running distance and speed such as Tiki-taka against a tactic such as Gegenpressing which involves high running distances and speed.

References

[1] Georgi Boshnakov, Tarak Kharrat, and Ian G. McHale. A bivariate weibull count model for forecasting association football scores. *International Journal of Forecasting*, 33(2):458 – 466, 2017.

[2] Julen Castellano, David Alvarez-Pastor, and Paul S. Bradley. Evaluation of research using computerised tracking systems (amisco and prozone) to analyze physical performance in elite soccer: A systematic review. *Sports Medicine*, 44(5):701–712, 2014.

[3] Luke Bornn Dan Cervone, Alexander D'Amour and Kirk Goldsberry. Pointwise: Predicting points and valuing decisions in real time with nba optical tracking data. *MIT Sloan Sports Analytics Conference Proceedings*, 2014.

[4] Ernesto Estrada and Naomichi Hatano. Communicability in complex networks. *Phys. Rev. E*, 77(3):036111, March 2008.

[5] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman and Hall/CRC, 1990.

[6] I. G. McHale and L. Szczepanski. A mixed effects model for identifying goal scoring ability of footballers. *Journal of the Royal Statistical Society, Series A*, 177(2):397–417, 2013.

[7] I. G. McHale and L. Szczepanski. Beyond completion rate: evaluating passing ability of footballers. *Journal of the Royal Statistical Society, Series A*, 178(4):513–533, 2015.

[8] Javier López Peña and Hugo Touchette. A network theory analysis of football strategies, 2012. Arxiv.1206.6904.

[9] E. Rampinini, A. J. Coutts, C. Castagna, R. Sassi, and F. M. Impellizzeri. Variation in top level soccer match performance. *Int. J. Sports Medicine*, 28(12):1018–1024, 2007.

[10] John W Tukey et al. Curves as parameters, and touch estimation. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

[11] Simon Wood and Fabian Scheipl. gamm4: Generalized additive mixed models using lme4 and mgcv, 2016. R package version 0.2-4.